Conservation Genet Resour
DOI 10.1007/s12686-016-0580-5

METHODS AND RESOURCES ARTICLE

Microsatellite discovery in an insular amphibian (*Grandisonia alternans*) with comments on cross-species utility and the accuracy of locus identification from unassembled Illumina data

Eleanor A. S. Adamson[1] · Anwesha Saha[1,2,3] · Simon T. Maddock[1,2,4] · Ronald A. Nussbaum[5] · David J. Gower[1] · Jeffrey W. Streicher[1]

Received: 9 March 2016 / Accepted: 25 July 2016

Abstract The Seychelles archipelago is unique among isolated oceanic islands because it features an endemic radiation of caecilian amphibians (Gymnophiona). In order to develop population genetics resources for this system, we identified microsatellite loci using unassembled Illumina MiSeq data generated from a genomic library of *Grandisonia alternans*, a species that occurs on multiple islands in the archipelago. Applying a recently described method (*PALFINDER*) we identified 8001 microsatellite loci that were potentially informative for population genetics analyses. Of these markers, we screened 60 loci using five individuals, directly sequenced several amplicons to confirm their identity, and then used eight loci to score allele sizes in 64 *G. alternans* individuals originating from five islands. A number of these individuals were sampled using non-lethal methods, demonstrating the efficacy of non-destructive molecular sampling in amphibian research. Although two loci satisfied our criteria as diploid, neutrally evolving loci with the statistical power to detect population structure, our success in identifying reliable loci was very low. Additionally, we discovered some issues with primer redundancy and differences between Illumina and Sanger sequences that suggest some Illumina-inferred loci are invalid. We investigated cross-species utility for eight loci and found most could be successfully amplified, sequenced and aligned across other species and genera of caecilians from the Seychelles. Thus, our study in part supported the validity of using *PALFINDER* with unassembled reads for microsatellite discovery within and across species, but importantly identified major limitations to applying this approach to small datasets (ca. 1 million reads) and loci with small tandem repeat sizes.

Keywords Caecilians · Cross amplification · Gymnophiona · *Hypogeophis* · Indotyphlidae · *Praslinia* · Seq-to-SSR approach · Simple sequence repeats

Eleanor A. S. Adamson and Anwesha Saha have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s12686-016-0580-5) contains supplementary material, which is available to authorized users.

✉ Jeffrey W. Streicher
j.streicher@nhm.ac.uk

[1] Department of Life Sciences, The Natural History Museum, London SW7 5BD, UK

[2] Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK

[3] Ashoka Trust for Research of Ecology and the Environment, Bangalore, India

[4] Department of Animal Management, Reaseheath College, Nantwich CW5 6DF, UK

[5] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

Introduction

Microsatellites or simple sequence repeats (SSR) are one of the most widely used tools in population genetics and conservation biology. As costs associated with high-throughput Next Generation DNA sequencing (NGS) have decreased, microsatellite discovery via genomic shotgun libraries (the Seq-to-SSR approach) has emerged as a reliable and economical methodology for non-model species (Abdelkrim et al. 2009; Allentoft et al. 2009; Castoe et al. 2012a). Several software packages have been developed to identify microsatellite loci this way including

Published online: 29 July 2016

MSATCOMMANDER (Faircloth 2008), *PALFINDER* (Castoe et al. 2012a), and *QDD* (Meglécz et al. 2010, 2014). Typically these methods use de novo assemblies or unassembled reads to identify SSR loci. The use of unassembled reads is a timesaving characteristic (Castoe et al. 2012a; Lance et al. 2013), however, several studies have cautioned against this approach when using short read data, such as from the Illumina platform. These studies have found that Illumina paired-end identified microsatellite loci can have an order of magnitude smaller success rate (in terms of amplification across multiple individuals and sequencing error) than those loci identified from longer 454 or PacBio reads (Drechsler et al. 2013; Wei et al. 2014). Despite this, the use of *PALFINDER* to identify SSR loci from short, unassembled Illumina reads is widespread (e.g. amphibians, Drechsler et al. 2013; Peterman et al. 2013; bivalves, O'Bryhim et al. 2012a; crustaceans, Stoutamore et al. 2012; fish, O'Bryhim et al. 2012b; Nunzuata et al. 2013; mammals, Barthelmess et al. 2013; reptiles, Castoe et al. 2012b). Interestingly, few studies have performed validation experiments using SSR loci identified from unassembled reads. Those that have (e.g. Mikheyev et al. 2010; Delmas et al. 2011; Castoe et al. 2012b) have typically used unassembled 454 reads, thus leaving the question of whether SSR loci can be reliably identified from shorter reads poorly addressed outside of a few studies (Drechsler et al. 2013; Wei et al. 2014).

In this study, we performed validation experiments (via direct sequencing and power analysis) on SSR loci identified with unassembled Illumina data from caecilians (Amphibia, Gymnophiona). Caecilians are limbless amphibians that are restricted mostly to parts of the wet tropics and typically possess a fossorial ecology (Gower and Wilkinson 2008). As such, they are generally rarely encountered and most species are poorly understood (Gower and Wilkinson 2005). Globally, there are ten families of caecilian (Wilkinson et al. 2011; Kamei et al. 2012). We focused on *Grandisonia alternans* (Stejneger, 1893), an indotyphlid species that is widely distributed across the Seychelles, an archipelago situated off the coast of east Africa in the Indian Ocean (Fig. 1). With both endemic frog and caecilian radiations the Seychelles is unique among isolated oceanic archipelagos (Nussbaum 1984). Given the exceptional evolutionary insights provided by species occurring on small islands (e.g. Warren et al. 2015), the generally high level of threat of extinction faced by such species (e.g. Daltry 2007), and the scarcity of knowledge of caecilian population genetics (e.g. Gower and Wilkinson 2005), we were motivated to develop population genetics tools for the endemic caecilians of the Seychelles.

To accomplish this we identified thousands of putative SSR loci from shotgun genomic sequencing, tested 60 loci to investigate intraspecific utility, generated Sanger sequences to examine sequence similarity to Illumina reads, identified eight loci with putatively desirable characteristics for population genetics and size-scored loci in multiple individuals of *G. alternans*. We also examined the cross-species utility of these eight loci in additional caecilian taxa and compared divergence in their flanking regions relative to *G. alternans* sequences. Based on our findings we describe (i) the discovery of six microsatellite loci that amplify across divergent caecilian species, including two loci that satisfy theoretical assumptions of selective neutrally in *G. alternans* putative populations, and (ii) several important limitations we encountered while using the Seq-to-SSR approach with unassembled Illumina reads.

Materials and methods

DNA isolation and microsatellite identification

We generated a shotgun genomic library from a pool of six barcoded Seychelles caecilian specimens, comprising representatives of five species: *Grandisonia alternans*, *G. larvata* (Ahl 1934), *G. sechellensis* (Boulenger 1911), *Hypogeophis brevis* Boulenger 1911 and *H. rostratus* (Cuvier 1829). Library preparation was performed by first extracting genomic DNA with a Qiagen DNeasy Blood and Tissue kit and then using a standard Illumina Nextera DNA kit to prepare the pooled sample for sequencing (see Lewis et al. 2014). Sequencing was performed using a 500-cycle v.2 reagent kit on an Illumina MiSeq at the Core Research laboratories of the Natural History Museum, London. We removed and trimmed low quality reads of unassembled paired-end data using Illumina software and the FASTX-Toolkit (available at http://hannonlab.cshl.edu/fastx_toolkit/links.html). Barcoded pooled data were de-multiplexed and only sequences from the target taxa *G. alternans* were retained, originating from a single *G. alternans* individual from Silhouette Island (UMMZ 192945). This single individual dataset was used in all downstream analysis and has been uploaded to the NCBI sequence read archive (SAMN04543719-20).

Potentially amplifiable loci (PAL) selection

We used *PALFINDER* software v 2.03 (Castoe et al. 2012a) to identify putative SSR loci from NGS output, run with default settings and applying the following criteria in Primer3 (Untergasser et al. 2012): (i) a minimum of eight tandem repeats for dinucleotide motifs, (ii) a minimum of six tandem repeats for trinucleotide motifs, and (iii) a minimum of six tandem repeats for tetranucleotide motifs.

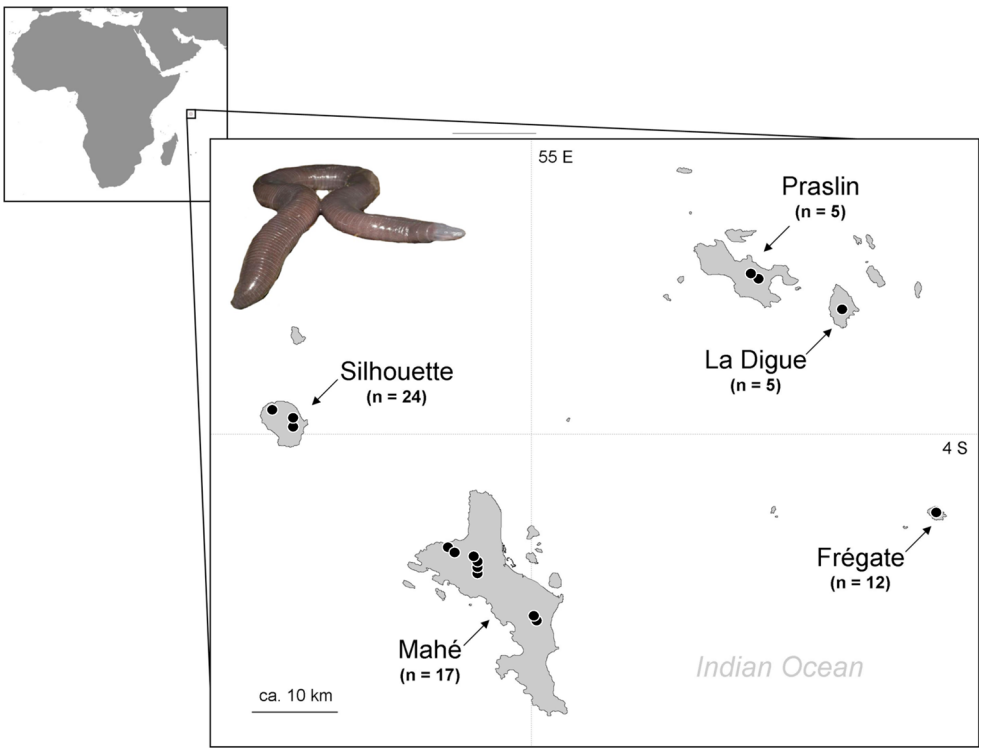

Fig. 1 Geographical distribution of sampling localities for individuals of *Grandisonia alternans* (Seychelles caecilians) that were screened for microsatellite variation in this study

From the *PALFINDER*/Primer3 output, a limited number of loci were chosen manually for further screening. All loci were predicted to contain simple microsatellite repeat regions and care was taken to retain loci that corresponded with predicted di-, tri-, and tetramer repeat motifs. The occurrence rate of unique priming regions in the unassembled reads varied considerably, and so primer sets were intentionally selected to represent a range of occurrence frequencies. This approach contrasts with the design of many previous studies, which intentionally avoided priming sites that occur frequently because they may be associated with repetitive elements. Thus, some selected primer sets corresponded to regions where both forward and reverse priming regions occurred up to 1000 times in the MiSeq reads, and some loci corresponded to where one or both priming regions occurred only once (e.g., Table 1). PALs were checked for homology against the NCBI's nucleotide database and retained only when they did not return high similarity to characterised sequence regions.

Taxonomic sampling and direct sequencing

The present study analysed 64 DNA samples of *G. alternans* (extracted from liver, buccal swab, annular clips and scales) from across five islands in the Seychelles (Mahé, La Digue, Praslin, Silhouette, and Frégate; Fig. 1; Supplementary Table). For non-lethal sampling we used the protocols described by Maddock et al. (2014). To address cross-species utility we also sampled five individuals representing other species of caecilian found in the Seychelles (*G. larvata* (Ahl, 1934), *G. sechellensis* (Boulenger, 1911), *Hypogeophis brevis* Boulenger, 1911, *H. rostratus* (Cuvier, 1829), and *Praslinia cooperi* Boulenger, 1909) and two taxa from peninsular India (*Gegeneophis ramaswamii* Taylor, 1964 and *Indotyphlus maharashtraensis* Giri, Wilkinson & Gower, 2004), from the sister clade to the Seychelles caecilians (e.g., San Mauro et al. 2014). DNA was extracted using a QIAGEN DNeasy Blood and Tissue Kit following manufacturer's instructions and diluted to concentrations of 20–100 ng/μL prior to PCR.

Microsatellite screening, fragment analysis, and scoring

Five *G. alternans* individuals from across the sampled geographical range of the species in the Seychelles archipelago were selected to investigate PCR amplification success and assess size variability in the PALs identified from the MiSeq data. All PCRs were conducted using either a Type-it Microsatellite PCR Kit or MyTaq^{TM} Red Mix with manufacturer's recommended reaction mix and cycling conditions. Initial trials tested a range of annealing temperatures for each locus (50–63 °C), all further PCRs used the optimum 60 °C annealing temperature. To visualise reaction success the amplified products were run on 3 % agarose gels (70 V, 80 min).

Table 1 Seychelles caecilian, *Grandisonia alternans*

Locus	Number of occurrences (forward primer)	Number of occurrences (reverse primer)	Number of occurrences (both primers)	Number of occurrences (PALs)
Galt 1	115	149	8	1
Galt 2	59	46	2	1
Galt 3	1	1	1	1
Galt 4	1	1	1	1
Galt 5	12	11	3	1
Galt 6	**38**	**83**	**1**	**1**
Galt 7	26	309	3	1
Galt 8	1	1	1	1
Galt 9*	1	1	1	1
Galt 10	196	231	6	1
Galt 11	185	13	2	1
Galt 12	184	196	2	1
Galt 13	**97**	**83**	**15**	**2**
Galt 14	**176**	**134**	**13**	**1**
Galt 15	**176**	**179**	**13**	**1**
Galt 16	**176**	**85**	**8**	**1**

PALs and corresponding primer sets that were identified from short read NGS sequencing and then successfully re-sequenced with Sanger technology. Numbers of occurrences refers to number of times that priming regions and PALs were present among the unassembled MiSeq data. Bolded text indicates loci that were duplicated (multiple primer sets actually amplified the same region) despite each corresponding PAL occurring only once in the *PALFINDER* output. Asterisk beside locus name indicates a PAL that did not have microsatellite regions when re-sequenced with Sanger technology

For every locus that showed clear gel bands of approximately the expected size range predicted by *PALFINDER* in each of the five test samples, one or two *G. alternans* individuals (including the original sample used for Illumina library preparation) were selected and the locus re-amplified, product cleaned by vacuum filtration and Sanger sequenced in both directions using Big Dye® Terminator Chemistry v3.1 on a 3730xl DNA Analyser by the DNA Sequencing Facility at the Natural History Museum, London. Chromatograms were compared to predicted PAL sequence and examined to confirm presence of microsatellite repeat region using Geneious R8 (Biomatters Ltd). Where microsatellite regions were confirmed, forward primers were synthesized with fluorescent dye labels prior to bulk screening individuals for size variation via Fragment Analysis (15 s injection) on a 3730xl DNA Analyser run with a Genescan 500 LIZ size standard.

Fragment analysis screening was conducted on all sampled individuals of *G. alternans*. PCRs used the protocols optimized in initial trials and were performed in 12.5 μL reactions with a heated lid, with or without multiplexing. Products were run on agarose to confirm PCR success prior to fragment analysis and product of single locus PCRs (non-multiplexed) were diluted and mixed together prior to fragment analysis so that each resulting *.fsa* file contained results for four loci with different fluorescent colour labels. Allele sizes were scored using the Geneious Microsatellite Plugin 1.4 using the local southern sizing method.

In addition to fragment analysis, loci that were consistently amplified were chosen to investigate cross-species utility. PCRs were conducted for single individuals of each species, and when a sharp band could be observed on agarose. Sanger sequencing was carried out as above. Directly sequenced PCR amplicons were submitted to GenBank under accession number KU739108-739133.

Data analysis

Microsatellite allele frequencies were compiled in Excel software v12.0 and input files and summary statistics were generated using CONVERT software v1.31 (Glaubitz 2004). To investigate the potential for the loci for future population genetic research, we tested their conformation to neutral expectations (i.e. Hardy–Weinberg Equilibrium: HWE, Linkage Equilibrium) using Arlequin software v3.5.1.3 (Excoffier and Lischer 2010). Because no a priori information exists as to what might constitute a panmictic population of *G. alternans*, samples were grouped by island for these analyses under the assumption that contemporary gene flow between islands for this species is highly unlikely.

Tests of statistical power were conducted to further investigate the potential of the loci to uncover genetic differentiation among populations, given the allelic

variation observed at each locus. Both the level of differentiation that it was possible to detect using current sampling and the number of samples required to detect relatively low differentiation (at or above F_{ST} values of 0.02) were investigated. Power analysis was performed using the program POWSIM (Ryman and Palm 2006) with each parameter set (N_e, t) employing 1000 replications.

Cross-species alignments were generated using Geneious R8 and further altered by hand. Approximate estimates of homology (percent base pair differences) were calculated after ends were trimmed and microsatellite regions excluded.

Results

PALs identified in Grandisonia alternans

After filtering results of pooled Illumina sequencing, 983,636 paired end shotgun reads were recovered from the genomic library of *G. alternans* individual UMMZ 192945. From these reads, *PALFINDER* identified 8001 microsatellite loci, but only 560 of these loci contained simple di-, tri-, or tetranucleotide motifs meeting our selection criteria (Fig. 2). Of these, 60 loci were selected for initial PCR, 20 of which were found to amplify consistently across five test individuals and were subsequently Sanger sequenced. Sixteen of the 20 PCR products produced readable sequence data that could be aligned at least partially with the PAL sequence, of which 15 loci contained microsatellite repeat regions (Table 1). The majority of loci that were successfully amplified and re-sequenced corresponded to loci for which both forward and reverse priming regions occurred together multiple times among all PAL sequences (Fig. 3).

Many of the PALs showed insertions when compared to Sanger sequence data, with most of these corresponding to duplications of near or adjacent regions (duplications ranged in size from 21 to 94 bps). Two PALs had substantial deletions (30, 150 bps) when compared to Sanger sequence generated from PCR. The majority (~90 %) of these indels did not occur beside microsatellite repeat regions (Fig. 4), suggesting that sequencing across potentially problematic repeat regions was not driving the observed sequencing errors. When taking these indels into account, close comparison between Sanger and PAL sequence across loci revealed that in two cases multiple primer sets were targeting the same genomic regions, reducing the number of independent microsatellites containing PALs to 11. These "duplicated" PALs corresponded to primer sets that usually, although not always, had relatively high occurrences of priming regions among the identified PALs (Table 1; Fig. 3).

Ultimately eight loci were selected that appear to amplify reliably and show di-allelic size variation; three with di-nucleotide repeat motifs, four with tri-nucleotide, and one with a tetranucleotide repeat motif (Galt 1-8; Table 2). Figure 4 illustrates the differences between PAL and Sanger sequences for these loci. Data were generated for 64 *G. alternans* individuals across these eight loci,

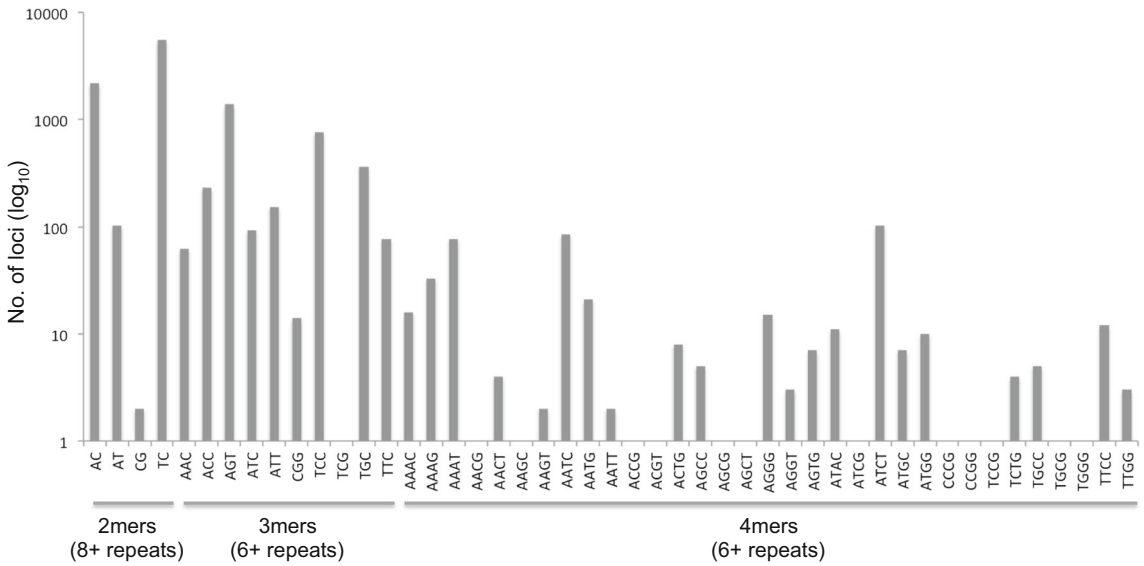

Fig. 2 Total number of PALs containing simple sequence repeat motifs that were identified from 983,636 unassembled Illumina paired-end reads sequenced from *Grandisonia alternans* (summarised by repeat motif)

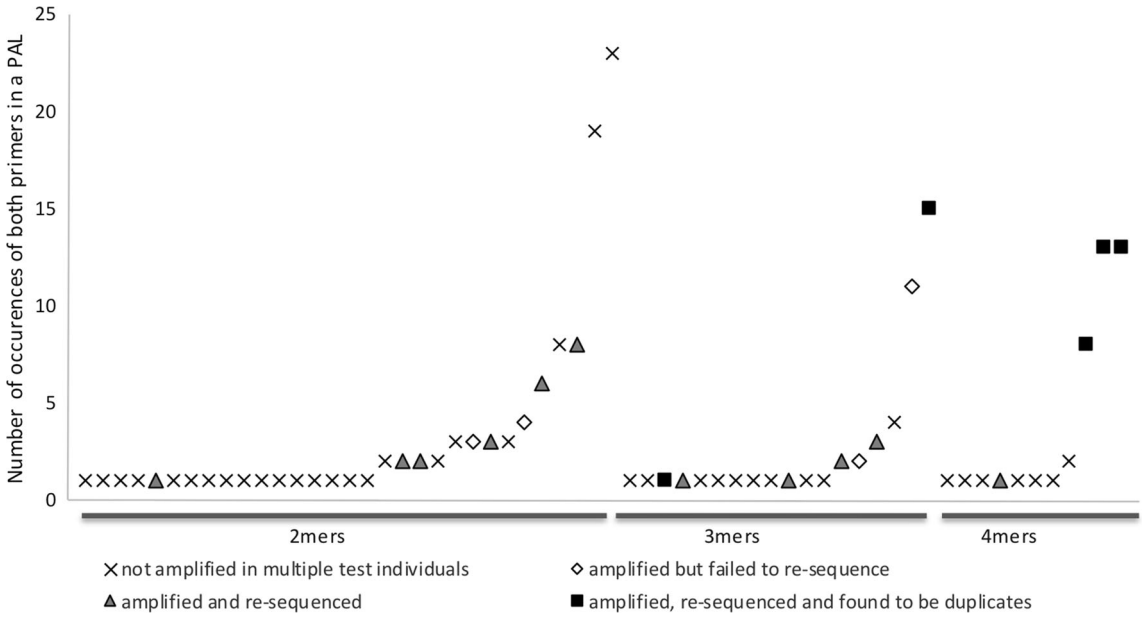

Fig. 3 Number of times that both forward and reverse primer sequences were present in PAL sequences for the 60 PALs trialed in this study. Loci were tested on five individuals, Sanger re-sequenced when possible, and then Sanger and PAL sequences were aligned to check for the presence of expected microsatellite repeats motifs and for duplication of amplified regions

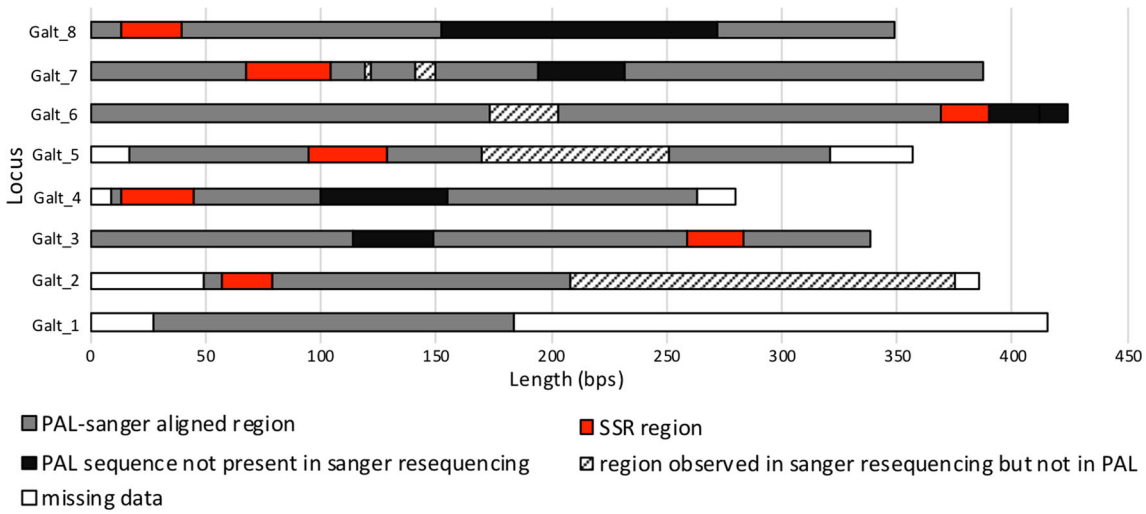

Fig. 4 Comparison between sequences predicted from shotgun sequencing data using *PALFINDER* (PALs) and sequences generated from Sanger sequencing with primers designed with *PALFINDER/PRIMER 3* for the eight focal loci. Seven of the PALs had indels when compared to Sanger sequencing data

however, we found evidence that the quality of these SSR loci (in terms of population genetics utility) was variable. Two loci (Galt 2 and 8) displayed behaviour consistent with informative SSRs. Four loci (Galt 1, 2, 6, and 7) appeared to be heterozygous but fixed across all individuals of *G. alternans* screened. One locus (Galt 3) possessed different allelic sizes, but we never observed a heterozygous individual. One locus (Galt 5) contained large numbers of stutter peaks and we were therefore unable to score it consistently.

Utility of SSR markers from Grandisonia alternans *for population genetics*

Only two PALs (Galt 4 and Galt 8) amplified well, could be reliably scored and showed size variation consistent with diploid SSR loci. Of the 64 individuals scored, one sample (BMNH 2005.1686) was not successfully genotyped at both these loci and was therefore excluded from further analysis. Within each of the five putative populations (grouped by island), no signature of linkage was detected

Table 2 Microsatellite/anonymous loci for *Grandisonia alternans*, annealing temp was 60 °C for all loci

Locus name	Primer sequence (5′–3′)	SSR motif	*G. alternans* microsatellite utility		Cross-species utility		
			Amplifies? Scorable? Variable? Diploid?	Number of alleles, size ranges (bp), and HWE conformation	Aligns with other taxa?	Other taxa contain SSR?	% Divergence
Galt 1	TCCTACCTTTGTTGTCTGGGC AAGAGAGAGACTGGATGGGGC	TC (n)	Yes (100 %) Yes Yes No	Every individual possessed three amplicons of sizes 261, 263, and 297.	No cross-amplification	N/A	N/A
Galt 2	TGTCTGTCGATGAGTCTCTGGC GCACAACATACACATTCATGCC	TCC (5)	Yes (98 %) No? N/A N/A	Indistinct range of peaks between 400–450	*H. brevis* *H. rostratus* *G. lavata* *G. sechellensis* *P. cooperi*	Yes Yes Yes Yes Yes	0.71 1.78 0.71 1.78 1.78
Galt 3	GTTGTGACCAGCAGGAGTCG GTGCTCCAGTCTTGCTTCCC	GGAA (4+)	Yes (90 %) Yes Yes N/A	Island specific size variants: Frégate, La Digue, Praslin = 333 Silhouette = 341 Mahé = 333 and 341 (always homozygous)	No cross-amplification	N/A	N/A
Galt 4	AAGAAGGTTGAATCCTCTCCCC TGTAACACCTACAATGAACATGGC	AG (11+)	Yes (98 %) Yes Yes Yes	12, 267–297, Yes (5/5)	*G. sechellensis*	Yes	0.00
Galt 5	GAGTGTGTAGAACAGGGTTGTCC GGACTTGAAACCATGGGACC	TC (11+)	Yes (95 %) No Yes? N/A	Too many stutter bands to score consistently.	*G. sechellensis*	Yes	5.28
Galt 6	CCCTAGAGATCACCCCTCCC CCCTTCCCAGTCTCCAGC	AGG (7+)	Yes (100 %) Yes Yes No	Every individual possessed three amplicons of sizes 433, 436, and 457.	*H. brevis* *H. rostratus* *G. sechellensis*	Yes Yes Yes	4.12 1.33 5.00
Galt 7	ATGACATTGCATCTGCGACC CAGAGGGTCAAGGTCTTCCC	AGG (4)	Yes (100 %) Yes Yes No	Every individual possessed three amplicons of sizes 355, 361, and 264.	*H. brevis* *H. rostratus* *G. sechellensis*	Yes (larger) Yes Probably	3.00 1.51 3.03
Galt 8	AGGCTGCAAAGGTGTTTTCC GGAGGATTAGAGCTTGCCCC	TGG (6+)	Yes (100 %) Yes Yes Yes	9, 261–285, Yes (3/5)	*H. brevis*	No	3.68

Various qualities of each locus are discussed further in text. Simple sequence repeat (SSR) motifs were confirmed via Sanger sequencing. Percent divergence was calculated using comparisons to *G. alternans* Sanger sequence data (excluding indels and repeat regions)

between loci (all exact p values >0.29). Data for locus Galt 4 conformed to HWE across all five islands (all p values >0.12). Significant departures from HWE were observed for locus Galt 8 among individuals drawn from the two largest of the five islands (Mahé $p < 0.0001$, Silhouette $p = 0.0415$), potentially indicating substructure at this geographical scale. When data for all islands was pooled and re-analysed significant departures from linkage and HWE were observed, consistent with the hypothesis that some level of genetic structuring is present for *G. alternans* across its distribution in the Seychelles archipelago.

Results of our power analysis of the two loci are presented in Fig. 5. Given five putative populations and levels of allelic variation observed from the 63 individuals

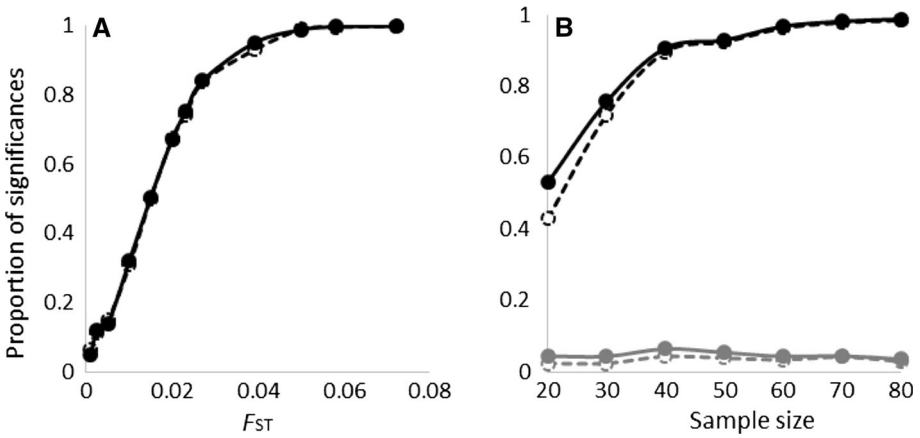

Fig. 5 Simulation estimates of power (proportion of significances) (*black*) and Type 1 error (*grey*) for the two microsatellite loci developed here (Galt 4 and Galt 8). *Closed circles* Chi squared analysis, *open circles* Fisher's test; **a** The data generated in this study from 63 individuals drawn for five putative populations have strong power (> 0.95) to detect significant differentiation at or above F_{ST} of ≈ 0.04, **b** results of simulations with expected divergence of $F_{ST} = 0.02$ and overall allelic variation observed in this study indicate the effect of sample size on the power and magnitude of Type 1 error. Using these loci only and drawing sample sizes of 40 individuals from each of two populations would be sufficiently powerful to detect F_{ST} differentiation as low as 0.02

surveyed here, the current two locus data set has the power to detect differentiation in the order of or above F_{ST} values of 0.04 if it is present (Fig. 5a). If sample sizes were increased to 40 per population, the two loci are likely to be useful in identifying differentiation as low as $F_{ST} = 0.02$ (Fig. 5b).

Cross-species amplifications

Six of the eight loci were successfully amplified and Sanger sequenced for other caecilian species, including members of the closely related Seychelles genera *Grandisonia*, *Hypogeophis* and *Praslinia*. When compared with *G. alternans* data, the maximum sequence divergence observed was 5.28 % (Table 2). SSRs were observed in different species for five of the six cross-amplified loci. None of the loci could be successfully amplified for the more distantly related Indian taxa using the *G. alternans* primer sets.

Discussion

Although we have identified two SSR loci suitable for describing population structure in the Seychelles caecilian, *Grandisonia alternans*, and six SSR loci that amplify across a range of other caecilian taxa, our investigation into the utility of using short read MiSeq Illumina sequence data to identify SSR loci suitable for population genetics was disappointing. We recognise issues associated with initial locus identification, with the quality of resulting markers, and with the amount of investment (both financial and in time) needed to use this approach. Although our approach differed in some respects from previous studies (using a relatively small NGS dataset, longer paired-end reads, considering a range of loci with respect to both tandem repeat type and frequency of occurrence in the NGS data set), the issues we identify are broadly applicable to validation of loci identified from shotgun datasets.

Using our data, we encountered two major limitations associated with the locus identification portion of the Seq-to-SSR method via *PALFINDER*: (i) multiple primer sets that amplified the same region and (ii) a PAL that lacked microsatellites. Both these limitations appear to be directly related to inaccuracies in the short read shotgun sequence data, resulting in PAL sequences that were substantially different from sequences obtained via direct Sanger re-sequencing. This problem is likely to be symptomatic of the type of data used, and we suspect that employing larger NGS effort (for example a full Illumina MiSeq lane for a single sample), while providing substantially more PALs, would not change the proportion of erroneous PALs encountered during locus validation.

In our case, of the 16 comparisons we performed between PAL and Sanger sequences generated for the same individual (Table 1), only two Sanger sequences could be aligned with the PAL sequence on which their priming regions had been designed without the introduction of large insertions (up to 120 bp) and or deletions (up to 167 bp). Indels relative to the *PALFINDER* sequences did not occur beside microsatellite sections, suggesting that sequencing errors are not related to problems sequencing across tandem repeat areas. Instead, they are likely related to poor concatenation of overlapping mate pairs, a hypothesis

supported by the insertions typically being short (non-microsatellite) duplications of adjacent sequence regions. A simple correlation test revealed no relationship between priming site occurrence rate and indel size in our alignments ($R^2 = 0.04$, $p = 0.47$ [two-tailed probability distribution]), indicating that indel patterns are not related to issues of PAL sequencing depth. However, there were two loci where we observed both insertions and deletions in Sanger and corresponding *PALFINDER* sequences (Galt 6 and 7; Fig. 4), which may indicate some further sequencing errors in at least one of our approaches.

The multiple primer sets that targeted the same genomic regions were not identified prior to Sanger sequencing because of indel differences in their PAL shotgun sequences. We recommend that future studies utilizing shotgun data should align PAL sequences from *PALFINDER* output allowing for the introduction of indels to help identify and eliminate the presence of such "duplicate loci". Although not a consistent pattern, it appears that priming regions of duplicated loci were generally represented in relatively high numbers among PALs (Table 1; Fig. 3), and were more prevalent in tri- and tetramer PALs, repeat motifs that are often considered better targets for locus design (Castoe et al. 2012a). It may be that selecting loci with moderate levels of primer occurrence as recommended by *PALFINDER* may minimize this problem. The lowest occurrences of priming sites among PALs, however, returned low success rates of useable loci, suggesting that there is likely to be a tradeoff between the risks of selecting duplicate loci and of choosing loci that can be amplified at all.

Our investigation into the Seq-to-SSR approach did identify two SSR loci that conformed to neutral expectations and had significant power to detect genetic structure in *G. alternans*, and so appear to represent ideal targets for further population genetic study of this species in the Seychelles. However, these were the minority of PAL loci originally identified using *PALFINDER* applying the recommended stringent criteria for identifying variable SSR loci with 2–4 tandem repeats (sensu Castoe et al. 2012a). The two loci that did pass our own criteria for population genetics microsatellites (i.e., reliable and repeatable amplification, size variability, diploidy, conformation to neutrality) were all that remained from the initial 60 primer sets that were trialed in the wet lab. The majority of PAL primer sets failed to amplify anything (Fig. 3), perhaps again due to errors in PAL sequence resulting in erroneous primer design. Of the eight loci that did make it through to the large scale screening of allelic variation, six either did not show patterns of variation consistent with diploid loci or were impossible to score consistently, rendering them unsuitable for most population genetics analyses that assume HWE. Excluding the many primer sets that we did not trial in the wet lab, our end result of 2/60 usable loci represents a disappointing success rate of only 3.3 %, a poor return for the time and money spent in the laboratory.

Perhaps more encouraging are our species cross-amplification results. Our study is not the first to demonstrate cross-species amplification of microsatellites in caecilians (Barratt et al. 2012). We did, however, observe several interesting patterns related to cross-species amplification among caecilians from the Seychelles. For example, in some microsatellite flanking regions the raw pairwise similarity between *G. alternans* and *Hypogeophis* spp. was smaller than between *G. alternans* and other *Grandisonia* species (*G. larvata* and *G. sechellensis*; Table 2). Although most of the species that were cross-amplified successfully were also present in the original pooled NGS run, we filtered sequences from other taxa out before designing and verifying the loci, so we consider our cross-amplification success to potentially reflect closeness of phylogenetic relationships rather than being an artefact of incorrect filtering of the pooled sample prior to locus design. These results are more likely related to non-monophyletic genus-level classification, homoplasy related to the rapid rate of SSR evolution, or differences in the rate of evolution among taxa; this will need to be revisited once a robust hypothesis of phylogenetic relationships is available for caecilian species from the Seychelles (for a recent perspective on this problem, see Maddock et al. 2016).

Also encouraging was our success in extracting and amplifying DNA from tissue collected with non-lethal sampling methods. The efficacy of non-lethal sampling for obtaining DNA suitable for Sanger sequencing has been demonstrated previously in caecilians (Maddock et al. 2014), however this study is the first to demonstrate the applicability of non-lethal sampling protocols for generating microsatellite data (see Prunier et al. (2012) for similar example in newts). This is encouraging in terms of conservation biology because it means that the deaths of large numbers of individuals of conservation concern will no longer be required to provide adequate sample sizes for population genetic studies.

Although we demonstrated an ability to use the Seq-to-SSR method to successfully identify loci that are useful for population genetics inference, the amount of time and financial investment in screening/size scoring was discouraging given the low success rate. Thus, although the identification of these potentially amplifiable loci may be rapid, the application of this method (via screening and size scoring) was, in our case, an inefficient approach. Based on these results, we support pursuing non-SSR loci methods (e.g. RADseq; Davey and Blaxter 2011), identifying SSRs from longer read sequence data (e.g. Drechsler et al. 2013; Wei et al. 2014), or restricting shotgun Seq-to-SSR searches to so called "Best PALs" (>6 repeats for 4–6 mers;

Castoe et al. 2012a) when researchers need to perform population genetics assessments using NGS with limited resources.

Acknowledgments We thank Julia Llewellyn-Hughes, Tom Pennance, Bonnie Webster and Fiona Allan (NHM, London) for help generating data. For additional practical assistance in obtaining material and generating and analyzing data we thank Wilna Accouche, Rachel Bristol, Nancy Bunbury, Lyndsay Chong Seng, Julia Day, Pat Dyal, Peter Foster, Erin Gleeson, Marc Jean-Baptiste, Rachunliu G. Kamei, Jim Labisko, Tanya Leibrick, Charles Morel, Greg Schneider, Bruno Simões, Martijn Timmermans, Gill Sparrow, and Mark Wilkinson. Funded in part by a EU-Marie Curie postdoctoral fellowship (to EASA), an NHM/UCL MRes Studentship (to AS), an NHM/UCL PhD Studentship (to STM), a SynTax grant (to Mark Wilkinson, Julia Day and DJG) and Darwin Initiative Grant 19-002. DJG and STM thank the Darwin Initiative grant partners in the Seychelles (Natural History Museum of the Seychelles, Seychelles Islands Foundation, Seychelles National Parks Authority, Island Conservation Society) and the UK. We would also like to thank Seychelles Bureau of Standards for permission to carry out fieldwork; Seychelles Department of Environment for permission to collect and export samples; Fregate Island Private for facilitating sampling on Frégate.

References

Abdelkrim J, Robertson B, Stanton JA, Gemmell N (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. Biotechniques 46:185–192

Allentoft M, Schuster SC, Holdaway R et al (2009) Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. Biotechniques 46:195–200

Barratt CD, Horsburgh GJ, Dawson DA et al (2012) Characterisation of nine microsatellite loci in the caecilian amphibian *Boulengerula uluguruensis* (Gymnophiona), and their cross-species utility in three congeneric species. Conserv Genet Resour 4:225–229

Barthelmess EL, Love CN, Jones KL, Lance SL (2013) Development of polymorphic microsatellite markers for the North American porcupine, *Erethizon dorsatum*, using paired-end Illumina sequencing. Conserv Genet Resour 5:925–927

Castoe TA, Poole AW, De Koning APJ et al (2012a) Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. PLoS One 7:e30953

Castoe TA, Streicher JW, Meik JM et al (2012b) Thousands of microsatellites from the venomous coralsnake (*Micrurus fulvius*) and variability of select loci across populations and related species. Mol Ecol Resour 12:1105–1113

Daltry JC (2007) An introduction to the herpetofauna of Antigua, Barbuda and Redonda, with some conservation recommendations. Appl Herpetol 4:97–130

Davey JW, Blaxter ML (2011) RADseq: next-generation population genetics. Brief Funct Genomics 9:416–423

Delmas CEL, Lhuillier E, Pornon A, Escaravage N (2011) Isolation and characterization of microsatellite loci in *Rhododendron ferrugineum* (Ericacae) using pyrosequencing technology. Am J Bot 98:e120–e122

Drechsler A, Geller D, Freund K et al (2013) What remains from a 454 run: estimation of success rates of microsatellite loci development in selected newt species (*Calotriton asper*, *Lissotriton helveticus*, and *Triturus cristatus*) and comparison with Illumina-based approaches. Ecol Evol 11:3947–3957

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567

Faircloth BC (2008) MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. Mol Ecol Resour 8:92–94

Glaubitz JC (2004) CONVERT: a user friendly program to reformat diploid genotypic data for commonly used population genetic software packages. Mol Ecol Notes 4:309–310

Gower DJ, Wilkinson M (2005) Conservation biology of caecilian amphibians. Conserv Biol 19:45–55

Gower DJ, Wilkinson M (2008) Caecilians (Gymnophiona). In: Stuart SN, Hoffmann M, Chanson JS, Cox NA, Berridge R, Ramani P, Young BE (eds) Threatened amphibians of the world. Lynx Ediciones, with IUCN—The World Conservation Union, Conservation International, and NatureServe, Barcelona, pp 19–20

Kamei RG, San Mauro D, Gower DJ, Van Bocxlaer I, Sherratt E, Thomas A, Babu S, Bossuyt F, Wilkinson M, Biju SD (2012) Discovery of a new family of amphibian from northeast India with ancient links to Africa. Proc R Soc B Biol Sci 279:2396–2401

Lance SL, Love CN, Nunziata SO et al (2013) 32 species validation of a new Illumina paired-end approach for the development of microsatellites. PLoS One 8:e81853

Lewis CJ, Maddock ST, Day JJ, Nussbaum RA, Morel C, Wilkinson M, Foster PG, Gower DJ (2014) Development of anonymous nuclear markers from Illumina paired-end data for Seychelles caecilian amphibians (Gymnophiona: Indotyphlidae). Conserv Genet Resour 6:289–291

Maddock ST, Lewis CJ, Wilkinson M, Day J, Morel C, Kouete M, Gower DJ (2014) Non-lethal DNA sampling for caecilian amphibians. Herpetol J 24:255–260

Maddock ST, Briscoe AG, Wilkinson M et al (2016) Next-generation mitogenomics: a comparison of approaches applied to caecilian amphibian phylogeny. PLoS One 11(6):e0156757

Meglécz E, Costedoat C, Dubut V, Gilles A, Malausa T, Pech N, Martin JF (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. Bioinformatics 26:403–404

Meglécz E, Pech N, Gilles A et al (2014) QDD version 3.1: a user friendly computer program for microsatellite selection and primer design revisited: experimental validation of variables determining genotyping success rate. Mol Ecol Resour 14:1302–1313

Mikheyev AS, Vo T, Wee B, Singer MC, Parmesan C (2010) Rapid microsatellite isolation from a butterfly by de novo transcriptome sequencing: performance and a comparison with AFLP-dervied distances. PLoS One 5:e11212

Nunzuata SO, Lance SL, Jones KL et al (2013) Development and characterization of twenty-three microsatellite markers for the freshwater minnow Santa Ana speckled dace (*Rhinichthys osculus* spp., Cyprinidae) using paired-end Illumina shotgun sequencing. Conserv Genet Resour 5:145–148

Nussbaum RA (1984) Amphibians of the Seychelles. In: Stoddart DR (ed) Biogeography and ecology of the Seychelles islands. Dr W Funk Publishers, Boston, pp 379–415

O'Bryhim JO, Chong JP, Lance SL et al (2012a) Development and characterization of sixteen microsatellite markers for the federally endangered species: *Leprodea leptodon* (Bivalvia:

Unionidae) using paired-end Illumina shotgun sequencing. Conserv Genet Resour 4:787–789

O'Bryhim JO, Somers C, Lance SL et al (2012b) Development and characterization of twenty-two novel microsatellite markers for the mountain whitefish, *Prosopium williamsoni* and cross-amplification in the round whitefish, *P. cylindraceum*, using paried Illumina shotgun sequencing. Conserv Genet Resour 5:89–91

Peterman WE, Pauley LR, Brocato ER et al (2013) Development and characterization of 22 microsatellite loci for the ringed salamander (*Ambystoma annulatum*) using paired-end Illumina shotgun sequencing. Conserv Genet Resour 5:993–995

Prunier J, Kaufmann B, Grolet O, Picard D et al (2012) Skin swabbing as a new efficient DNA sampling technique in amphibians, and 14 new microsatellite markers in the alpine newt (*Ichthyosaura alpestris*). Mol Ecol Resour 12:524–531

Ryman N, Palm S (2006) POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. Mol Ecol 6:600–602

San Mauro D, Gower DJ, Müller H, Loader SP, Zardoya R, Nussbaum RA, Wilkinson M (2014) Life-history evolution and mitogenomic phylogeny of caecilian amphibians. Mol Phylogenet Evol 73:177–189

Stoutamore JL, Love CN, Lance SL (2012) Development of polymorphic microsatellite markers for blue king crab (*Paralithodes platypus*). Conserv Genet Resour 4:897–899

Untergasser A, Cutcutache I, Koressaar T et al (2012) Primer3-new capabilities and interfaces. Nucl Acids Res 40:e115

Warren BH, Simberloff D, Ricklefs RE et al (2015) Islands as model systems in ecology and evolution: prospects fifty years after MacArthur-Wilson. Ecol Lett 18:200–217

Wei N, Bemmels JB, Dick CW (2014) The effects of read length, quality and quantity on microsatellite discovery and primer development: from Illumina to PacBio. Mol Ecol Resour 14:953–965

Wilkinson M, San Mauro D, Sherratt E, Gower DJ (2011) A nine-family classification of caecilians (Amphibia: Gymnophiona). Zootaxa 2874:41–64